Scalable Visual Comparison of Biological Trees and Sequences

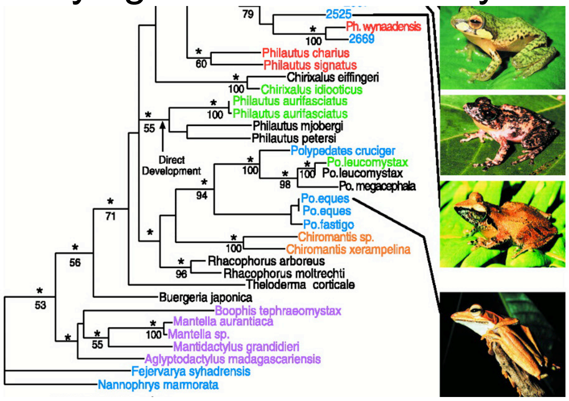
Tamara Munzner
University of British Columbia
Department of Computer Science

Imager

Outline

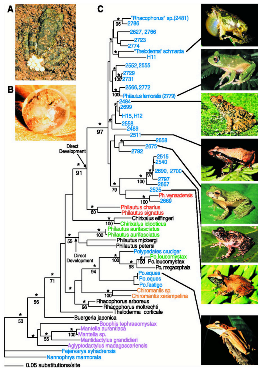
- TreeJuxtaposer
  - tree comparison
- Accordion Drawing
  - information visualization technique
- SequenceJuxtaposer
  - sequence comparison
- PRISAD
  - generic accordion drawing framework
- Evaluation
  - comparing AD to pan/zoom, with/without overview

Phylogenetic/Evolutionary Tree



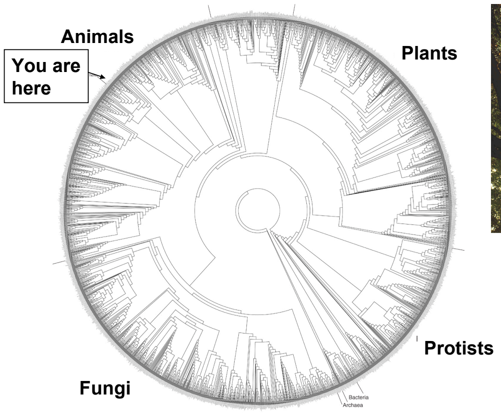
M Meegaskumbura et al., Science 298:379 (2002)

Common Dataset Size Today



M Meegaskumbura et al., Science 298:379 (2002)

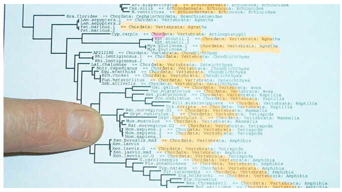
Future Goal: 10M Node Tree of Life



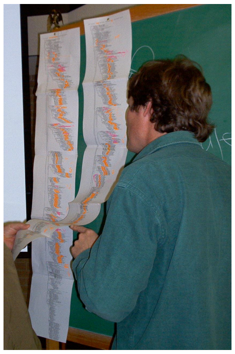
David Hillis, Science 300:1687 (2003)

Paper Comparison: Multiple Trees

focus

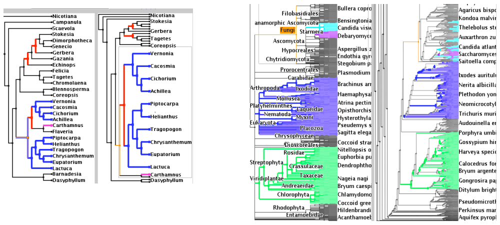


context



TreeJuxtaposer

- side by side comparison of evolutionary trees
  - [video]
  - software downloadable from http://olduvai.sf.net/tj



[TreeJuxtaposer: Scalable Tree Comparison using Focus+Context with Guaranteed Visibility. Tamara Munzner, François Guimbretière, Serdar Tasiran, Li Zhang, Yunhong Zhou. Proc SIGGRAPH 2003]

Related Work: Tree Browsing

- general
  - Cone Trees [Robertson et al 91]
  - Hyperbolic Trees [Lamping 94]
  - H3 [Munzner 97]
  - Hierarchical Clustering Explorer [Seo & Shneiderman 02]
  - SpaceTree [Plaisant et al 02]
  - DOI Tree [Card and Nation 02]
- phylogenetic trees
  - TreeWiz [Rost and Bornberg-Bauer 02]
  - TaxonTree [Lee et al 04]

Related Work: Comparison

- tree comparison
  - RF distance [Robinson and Foulds 81]
  - perfect node matching [Day 85]
- visual tree comparison
  - creation/deletion only [Chi and Card 99]
  - leaves only [Graham and Kennedy 01]
- subsequent work
  - DoubleTree [Parr et al 04]

TJ Contributions

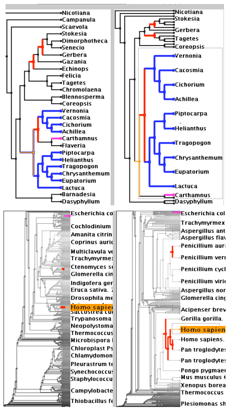
- first interactive tree comparison system
  - automatic structural difference computation
- scalable to large datasets
  - 250,000 to 500,000 total nodes
  - all preprocessing subquadratic
  - all realtime rendering sublinear
    - items to render >> number of available pixels
- scalable to large displays (4000 x 2000)
- introduced accordion drawing

Outline

- TreeJuxtaposer
  - tree comparison
- Accordion Drawing
  - information visualization technique
- SequenceJuxtaposer
  - sequence comparison
- PRISAD
  - generic accordion drawing framework
- Evaluation
  - comparing AD to pan/zoom, with/without overview

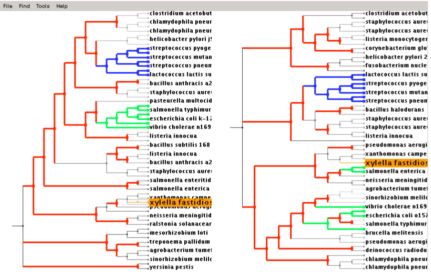
Accordion Drawing

- rubber-sheet navigation
  - stretch out part of surface, the rest squishes
  - borders nailed down
  - Focus+Context technique
    - integrated overview, details
  - old idea
    - [Sarkar et al 93], [Robertson et al 91]
- guaranteed visibility
  - marks always visible
  - important for scalability
  - new idea
    - [Munzner et al 03]



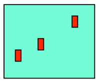
Guaranteed Visibility

- marks are always visible
  - regions of interest shown with color highlights
  - search results, structural differences, user specified
- easy with small datasets



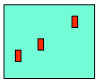
Guaranteed Visibility Challenges

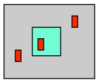
- hard with larger datasets
- reasons a mark could be invisible



Guaranteed Visibility Challenges

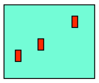
- hard with larger datasets
- reasons a mark could be invisible
  - outside the window
    - AD solution: constrained navigation

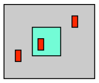


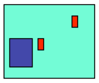


Guaranteed Visibility Challenges

- hard with larger datasets
- reasons a mark could be invisible
  - outside the window
    - AD solution: constrained navigation
  - underneath other marks
    - AD solution: avoid 3D

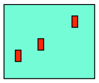


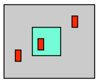


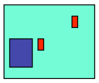


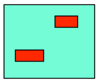
Guaranteed Visibility Challenges

- hard with larger datasets
- reasons a mark could be invisible
  - outside the window
    - AD solution: constrained navigation
  - underneath other marks
    - AD solution: avoid 3D
  - smaller than a pixel
    - AD solution: smart culling



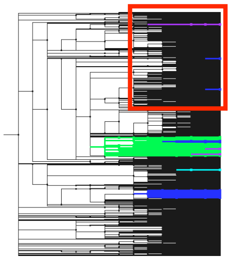




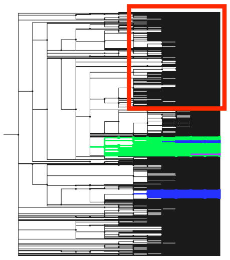


Guaranteed Visibility: Small Items

- Naïve culling may not draw all marked items



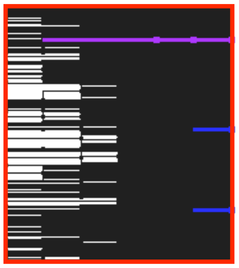
Guaranteed visibility of marks



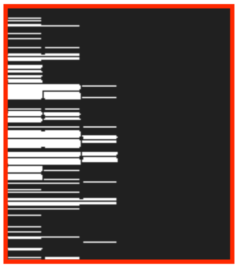
No guaranteed visibility

Guaranteed Visibility: Small Items

- Naïve culling may not draw all marked items



Guaranteed visibility of marks



No guaranteed visibility

Guaranteed Visibility Rationale

- relief from exhaustive exploration
  - missed marks lead to false conclusions
  - hard to determine completion
  - tedious, error-prone
- compelling reason for Focus+Context
  - controversy: does distortion help or hurt?
  - strong rationale for comparison
- infrastructure needed for efficient computation

Related Work

- multiscale zooming
  - Pad++ [Bederson and Hollan 94]
- multiscale visibility
  - space-scale diagrams [Furnas & Bederson 95]
  - effective view navigation [Furnas 97]
  - critical zones [Jul and Furnas 98]

Outline

- TreeJuxtaposer
  - tree comparison
- Accordion Drawing
  - information visualization technique
- SequenceJuxtaposer
  - sequence comparison
- PRISAD
  - generic accordion drawing framework
- Evaluation
  - comparing AD to pan/zoom, with/without overview

Genomic Sequences

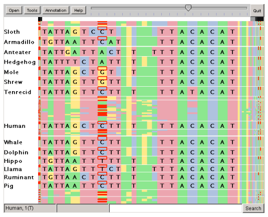
- multiple aligned sequences of DNA
- investigate benefits of accordion drawing
  - showing multiple focus areas in context
  - smooth transitions between states
  - guaranteed visibility for globally visible landmarks
- now commonly browsed with web apps
  - zoom and pan with abrupt jumps

Related Work

- web based, database driven, multiple tracks
  - Ensembl [Hubbard 02]
  - UCSC Genome Browser [Kent 02]
  - NCBI [Wheeler 02]
- client side approaches
  - Artemis [Rutherford et al 00]
  - BARD [Spell et al 03]
  - PhyloVISTA [Shah et al 03]

SequenceJuxtaposer

- side by side comparison of multiple aligned gene sequences
- [video], software downloadable from http://olduvai.sf.net/sj

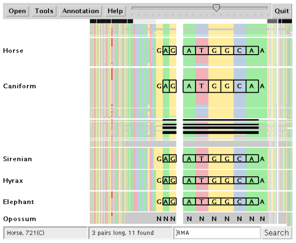

Open
Tools
Annotation
Help
Quit
Sloth TATTAGTCCT TTACACAT
Armadillo TGTTAATTCAT TTACACAT
Anteater TATTGATTACT T TTTACACAT
Hedgehog TATTTTCTATT T TTACACAT
Mole TATTAGCTGT T TTACACAT
Shrew TATTAGTTGTT TTACACAT
Tenrecid TATTAGTTCTT T TTATACAT
Human TATTAGCTCTT T TTACACAT
Whale TATTAGTTCTT TTACACAT
Dolphin TATTAGTTCTT TTACACAT
Hippo TGTTAATTTTT T TTACACAT
Llama TATTAGTTTCT T TTTACACAT
Ruminant TGTTAACTCTT T TTACACAT
Pig TATTAATTCTT T TTACACAT
Search

Human, 1(T)

[SequenceJuxtaposer: Fluid Navigation For Large-Scale Sequence Comparison In Context. James Slack, Kristian Hildebrand, Tamara Munzner, and Katherine St. John. Proc. German Conference on Bioinformatics 2004]

Searching

- search for motifs
  - protein/codon search
  - regular expressions supported
- results marked with guaranteed visibility



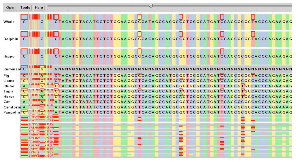
Horse, 721(C)

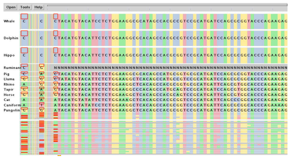
3 pairs long, 11 found

RMA

Differences

- explore differences between aligned pairs
  - slider controls difference threshold in realtime
  - standard difference algorithm, not novel
- results marked with guaranteed visibility





SJ Contributions

- fluid tree comparison system
  - showing multiple focus areas in context
  - guaranteed visibility of marked areas
    - thresholded differences, search results
- scalable to large datasets
  - 2M nucleotides
  - all realtime rendering sublinear

Outline

- TreeJuxtaposer
  - tree comparison
- Accordion Drawing
  - information visualization technique
- SequenceJuxtaposer
  - sequence comparison
- PRISAD
  - generic accordion drawing framework
- Evaluation
  - comparing AD to pan/zoom, with/without overview

Scaling Up: TJC/TJC-Q

- TJC: 15M nodes
  - no quadtree
  - picking with new hardware feature
    - requires HW multiple render target support
- TJC-Q: 5M nodes
  - lightweight quadtree for picking support
- both support tree browsing only
  - no comparison data structures

[Scalable, Robust Visualization of Large Trees Dale Beermann, Tamara Munzner, Greg Humphreys. Proc. EuroVis 2005]

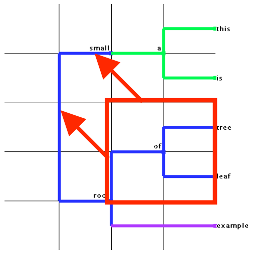
Generic Infrastructure: PRISAD

- generic AD infrastructure
  - PRITree is TreeJuxtaposer using PRISAD
  - PRISeq is SequenceJuxtaposer using PRISAD
- efficiency
  - faster rendering: minimize overdrawing
  - smaller memory footprint
- correctness
  - rendering with no gaps: eliminate overculling

[Partitioned Rendering Infrastructure for Scalable Accordion Drawing.
James Slack, Kristian Hildebrand, and Tamara Munzner.
Proc. InfoVis 2005
extended version: Information Visualization, to appear]

Navigation

- generic navigation infrastructure
  - application independent
  - uses deformable grid
  - split lines
    - grid lines define object boundaries
  - horizontal and vertical separate
    - independently movable

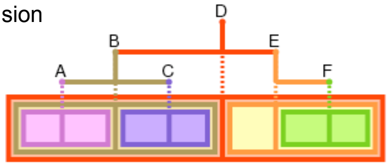


Split Line Hierarchy

- data structure supports navigation, picking, drawing
- two interpretations
  - linear ordering

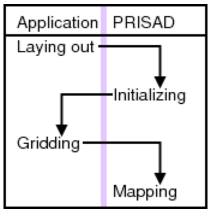
A	B	C	D	E	F

  - hierarchical subdivision

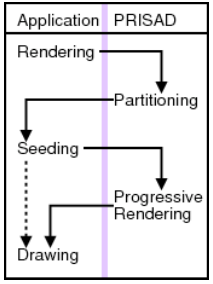


PRISAD Architecture

world-space discretization
- preprocessing
  - initializing data structures
  - placing geometry

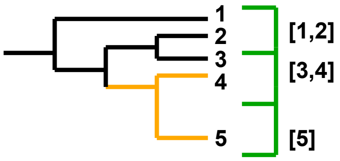


screen-space rendering
- frame updating
  - analyzing navigation state
  - drawing geometry



Partitioning

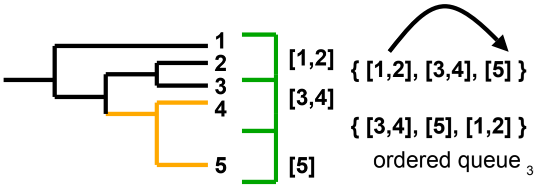
- partition object set into bite-sized ranges
  - using current split line screen-space positions
    - required for every frame
  - subdivision stops if region smaller than 1 pixel
    - or if range contains only 1 object



{ [1,2], [3,4], [5] }
Queue of ranges

Seeding

- reordering range queue result from partition
  - marked regions get priority in queue
    - drawn first to provide landmarks



{ [1,2], [3,4], [5] }

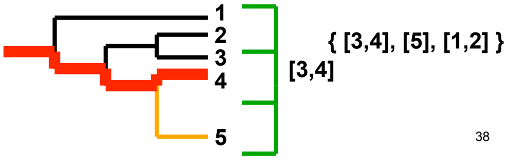
{ [3,4], [5], [1,2] }
ordered queue

Drawing Single Range

- each enqueued object range drawn according to application geometry
  - selection for trees
  - aggregation for sequences

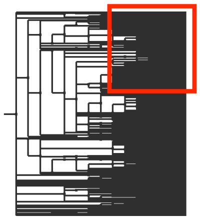
PRITree Range Drawing

- select suitable leaf in each range
- draw path from leaf to the root
  - ascent-based tree drawing
  - efficiency: minimize overdrawing
    - only draw one path per range

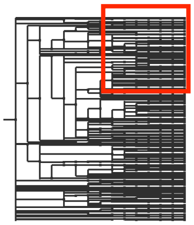


Rendering Dense Regions

- correctness: eliminate overculling
  - bad leaf choices would result in misleading gaps
- efficiency: maximize partition size to reduce rendering
  - too much reduction would result in gaps



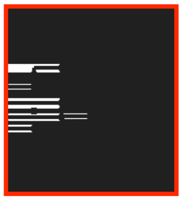
Intended rendering



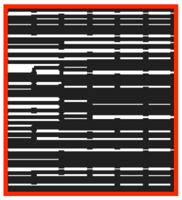
Partition size too big

Rendering Dense Regions

- correctness: eliminate overculling
  - bad leaf choices would result in misleading gaps
- efficiency: maximize partition size to reduce rendering
  - too much reduction would result in gaps



Intended rendering

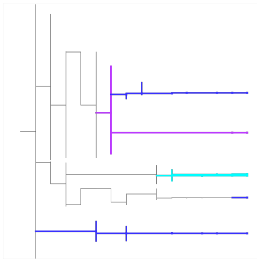


Partition size too big

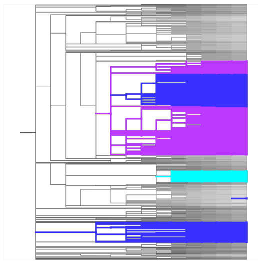
PRITree Skeleton

- guaranteed visibility of marked subtrees during progressive rendering

first frame: one path
per marked group

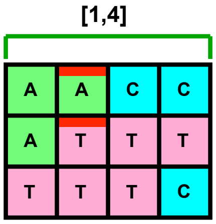


full scene:
entire marked subtrees

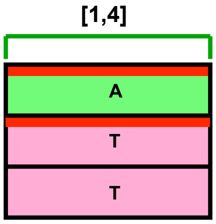


PRISeq Range Drawing: Aggregation

- aggregate range to select box color for each sequence
  - random select to break ties



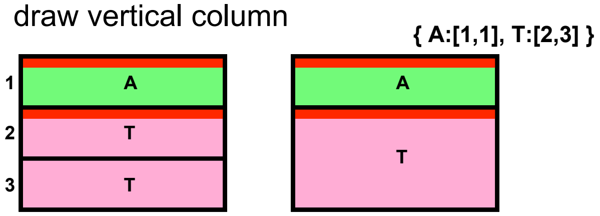
A	A	C	C
A	T	T	T
T	T	T	C



A
T
T

PRISeq Range Drawing

- collect identical nucleotides in column
  - – form single box to represent identical objects
    - attach to split line hierarchy cache
    - lazy evaluation
- draw vertical column



{ A:[1,1], T:[2,3] }

1 A

2 T

3 T

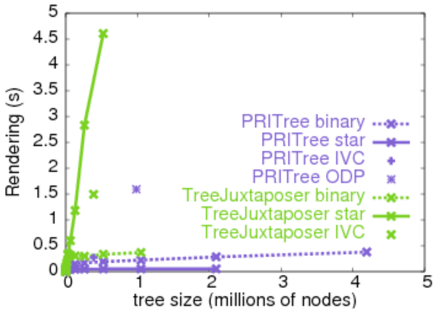
A

T

PRITree Rendering Time Performance

TreeJuxtaposer renders **all** nodes for star trees

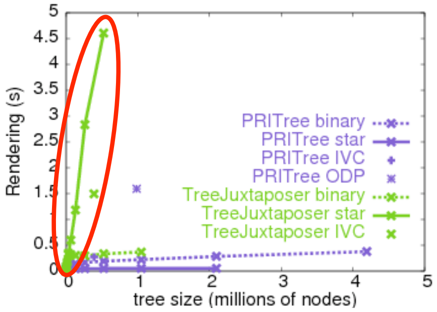
- branching factor k leads to O(k) performance



PRITree Rendering Time Performance

TreeJuxtaposer renders **all** nodes for star trees

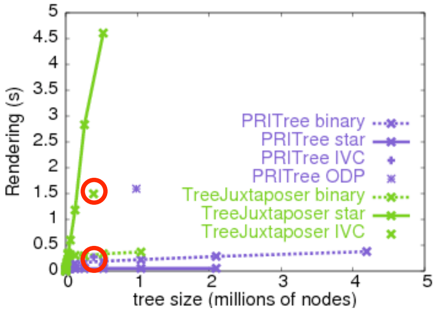
- branching factor k leads to O(k) performance



PRITree Rendering Time Performance

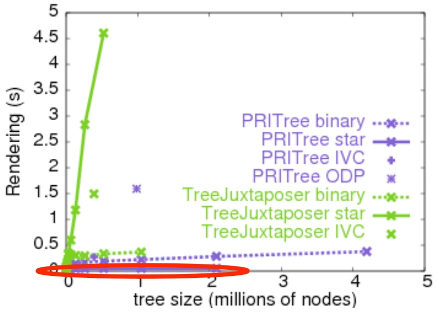
InfoVis 2003 Contest dataset

- 5x rendering speedup

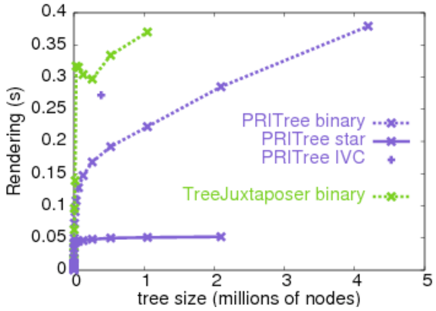


PRITree Rendering Time Performance

a closer look at the fastest rendering times



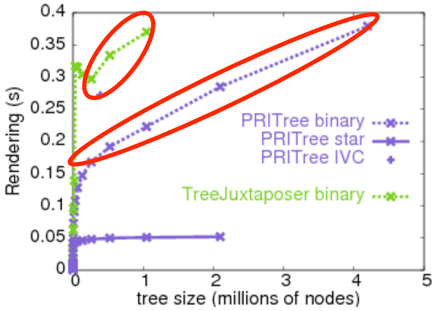
PRITree Rendering Time Performance



Detailed Rendering Time Performance

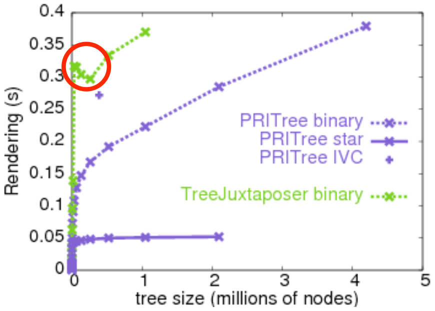
PRITree handles 4 million nodes in under 0.4 seconds

- TreeJuxtaposer takes twice as long to render 1 million nodes



Detailed Rendering Time Performance

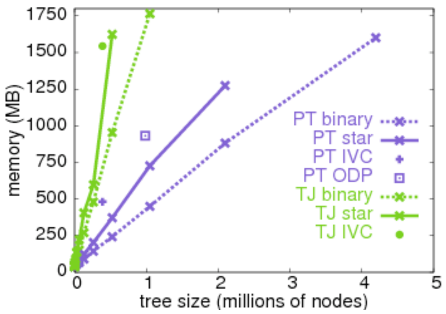
TreeJuxtaposer valley from overculling



Memory Performance

linear memory usage for both applications

- 4-5x more efficient for synthetic datasets



Performance Comparison

- PRITree vs. TreeJuxtaposer
  - detailed benchmarks against identical TJ functionality
    - 5x faster, 8x smaller footprint
    - handles over 4M node trees
- PRISeq vs. SequenceJuxtaposer
  - 15x faster rendering, 20x smaller memory size
  - 44 species * 17K nucleotides = 770K items
  - 6400 species * 6400 nucleotides = 40M items

PRISAD Contributions

- infrastructure for efficient, correct, and generic accordion drawing
- efficient and correct rendering
  - screen-space partitioning tightly bounds overdrawing and eliminates overculling
- first generic AD infrastructure
  - PRITree renders 5x faster than TJ
  - PRISeq renders 20x larger datasets than SJ
- future work
  - editing support

Outline

- TreeJuxtaposer
  - tree comparison
- Accordion Drawing
  - information visualization technique
- SequenceJuxtaposer
  - sequence comparison
- PRISAD
  - generic accordion drawing framework
- Evaluation
  - comparing AD to pan/zoom, with/without overview

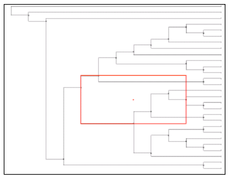
Evaluation

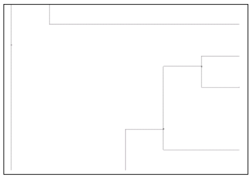
- evaluate RSN navigation technique
  - compare to conventional pan/zoom
- clarify utility of overviews for navigation
  - why add overview to F+C?
    - Need evidence to support or refute common InfoVis assumption regarding usefulness of overviews

[An Evaluation of Pan & Zoom and Rubber Sheet Navigation with and without an Overview. Dmitry Nekrasovski, Adam Bodnar, Joanna McGrenere, François Guimbretière, and Tamara Munzner. Proc. SIGCHI 06.

Conventional Pan & Zoom (PZN)

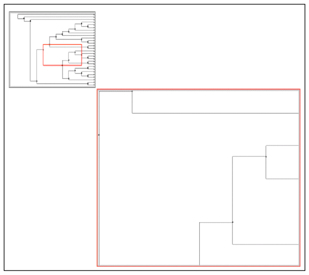
- navigation via panning (translation) and zooming (uniform scale changes)
- easy to lose context and become lost


Selecting region to zoom


Zooming result

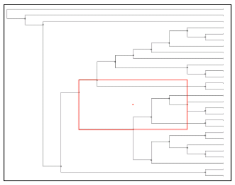
Overviews

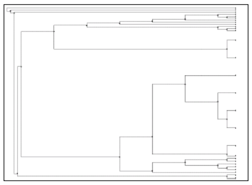
- separate global view of the dataset
- maintain contextual awareness
- force attention split between views



Rubber Sheet Navigation (RSN)

- Focus + Context technique
- stretching and squishing rubber sheet metaphor
- maintain contextual awareness in single view


Selecting region to zoom


Zooming result

Previous Findings Mixed

- mixed results for navigation and overviews
- speed: F+C faster than PZN

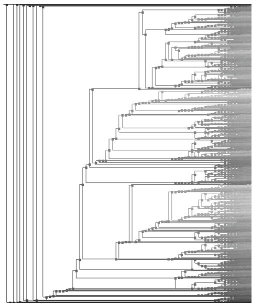
  [Schaffer et al., 1996; Gutwin and Skopik, 2003]
- accuracy: PZN more accurate than F+C

  [Hornbaek and Frokjaer, 2001; Gutwin and Fedak, 2004]
- preference: Overviews generally preferred

  [Beard and Walker, 1990; Plaisant et al., 2002]

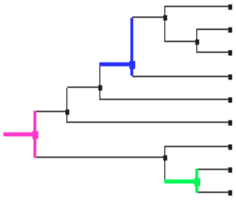
Dataset

- Motivating domain: evolutionary biology
  - large datasets, clear tasks
  - require understanding topological structure at different places and scales
- 5,918 node binary tree
  - Leaves are species, internal nodes are ancestors



Task

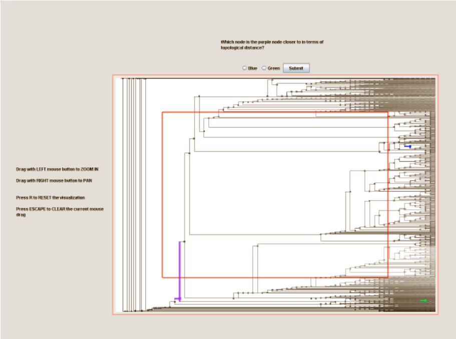
- Generalized version requiring no specialized knowledge of evolutionary trees (no labels)
- Compare topological distance between marked nodes
- Requires multiple navigation actions to complete
- Several instances isomorphic in difficulty



Experiment Interfaces

- Common visual representation and interaction model
  - Lacking in majority of previous evaluations
- Common set of navigation actions
- Guarantee visibility of areas of interest

RSN



Which node is the purple node closer to in terms of topological distance?

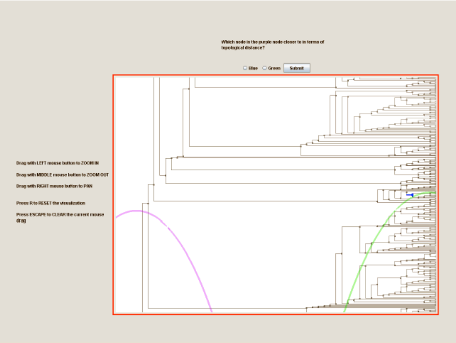
Drag with LEFT mouse button to ZOOM IN

Drag with RIGHT mouse button to PAN

Press R to RESET the visualization

Press ESCAPE to CLEAR the current mouse drag

PZN



Which node is the purple node closer to in terms of topological distance?

Drag with LEFT mouse button to ZOOM IN

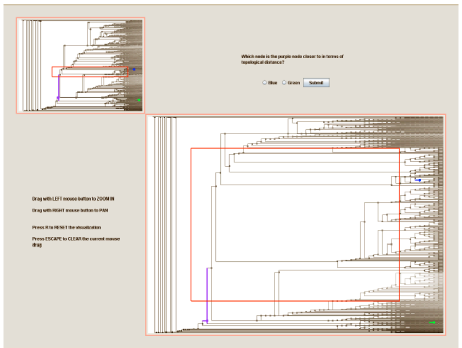
Drag with MIDDLE mouse button to ZOOM OUT

Drag with RIGHT mouse button to PAN

Press R to RESET the visualization

Press ESCAPE to CLEAR the current mouse drag

RSN + Overview



Which node is the purple node closer to in terms of topological distance?

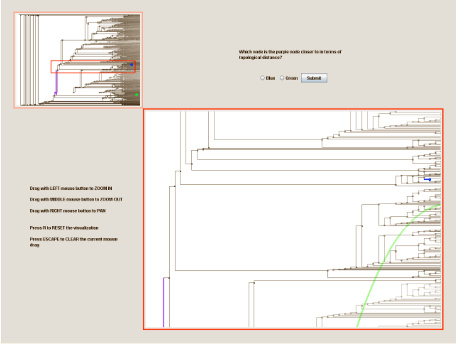
Drag with LEFT mouse button to ZOOM IN

Drag with RIGHT mouse button to PAN

Press R to RESET the visualization

Press ESCAPE to CLEAR the current mouse drag

PZN + Overview



Which node is the purple node closer to in terms of topological distance?

Drag with LEFT mouse button to ZOOM IN

Drag with MIDDLE mouse button to ZOOM OUT

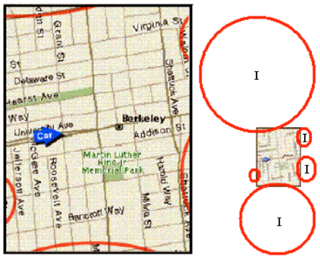
Drag with RIGHT mouse button to PAN

Press R to RESET the visualization

Press ESCAPE to CLEAR the current mouse drag

Guaranteed Visibility

- PZN
  – Implemented in PZN similarly to Halo [Baudisch et al., 2003]
- RSN
  – Implicit as areas of interest compressed along bounds of display
- Sub-pixel marked regions always drawn using PRISAD framework [Slack et al., 2005]



Hypotheses

H1 - RSN performs better than PZN independent of overview presence

H2 - For RSN, presence of overview does not result in better performance

H3 - For PZN, presence of overview results in better performance

Design

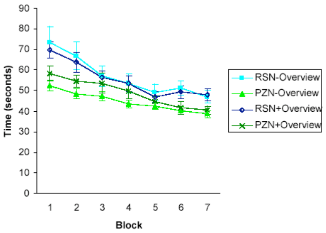
- 2 (navigation, between) x 2 (presence of overview, between) x 7 (blocks, within)
- Each block contained 5 randomized trials
- 40 subjects, each randomly assigned to each interface

Procedure and Measures

- Training protocols used to train subjects in effective strategies to solve task
- Subjects completed 35 trials (7 blocks x 5 trials), each isomorphic in difficulty
- Completion time, navigation actions, resets, errors, and subjective NASA-TLX workload

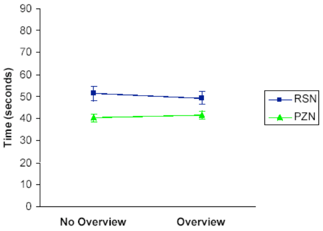
Results - Navigation

- PZN outperformed RSN ($p < 0.001$)
- Learning effect shows performance plateau
- Subjects using PZN performed fewer navigation actions and fewer resets
- Subjects using PZN reported less mental demand ($p < 0.05$)



Results – Presence of Overview

- No effect on any performance measure
- Subjects using overviews reported less physical demand and more enjoyment ($p < 0.05$)



Summary of Results

H 1 - RSN performs better than PZN independent of overview presence

- No – PZN outperformed RSN

H 2 - For RSN, presence of overview does not result in better performance

- Yes – No effect of overview on performance

H 3 - For PZN, presence of overview results in better performance

- No – No effect of overview on performance

Discussion – Navigation

- Performance differences cannot be ascribed to unfamiliarity with the techniques
- Design guidelines for PZN extensively studied, but not so for F+C or RSN

Discussion – Overviews

- Overviews for PZN and RSN:
  - No performance benefits
  - Preference for overview
- Overview may act as *cognitive cushion*
  - Provide subjective but not performance benefits
- Guaranteed visibility may provide same benefits as overviews

Evaluation Conclusions

- First evaluation comparing PZN and RSN techniques with and without an overview
- Performance:
  - PZN faster and more accurate than RSN
- Preference:
  - Overviews preferred, but no performance benefits

Other Projects

- Focus+Context evaluation
  - low-level visual search and visual memory
- graph drawing
  - TopoLayout: multi-level decomposition and layout using topological features
- dimensionality reduction
  - MDSteer: progressive and steerable MDS
- papers, talks, videos available from http://www.cs.ubc.ca/~tmm